Scene-adaptive and Region-aware Multi-modal Prompt for Open Vocabulary Object Detection

Xiaowei Zhao[1] * Xianglong Liu[1,2,3†] Duorui Wang[1] * Yajun Gao[1] Zhide Liu[1]
[1]State Key Laboratory of Complex & Critical Software Environment, Beihang University
[2]Zhongguancun Laboratory [3]Institute of data space, Heifei Comprehensive National Science Center

Abstract

Open Vocabulary Object Detection (OVD) aims to detect objects from novel classes described by text inputs based on the generalization ability of trained classes. Existing methods mainly focus on transferring knowledge from large Vision and Language models (VLM) to detectors through knowledge distillation. However, these approaches show weak ability in adapting to diverse classes and aligning between the image-level pre-training and region-level detection, thereby impeding effective knowledge transfer. Motivated by the prompt tuning, we propose scene-adaptive and region-aware multi-modal prompts to address these issues by effectively adapting class-aware knowledge from VLM to the detector at the region level. Specifically, to enhance the adaptability to diverse classes, we design a scene-adaptive prompt generator from a scene perspective to consider both the commonality and diversity of the class distributions, and formulate a novel selection mechanism to facilitate the acquisition of common knowledge across all classes and specific insights relevant to each scene. Meanwhile, to bridge the gap between the pre-trained model and the detector, we present a region-aware multi-modal alignment module, which employs the region prompt to incorporate the positional information for feature distillation and integrates textual prompts to align visual and linguistic representations. Extensive experimental results demonstrate that the proposed method significantly outperforms the state-of-the-art models on the OV-COCO and OV-LVIS datasets, surpassing the current method by 3.0% mAP and 4.6% AP_r.

1. Introduction

Recognizing and localizing visual objects in images [20, 26, 28] is a fundamental problem in computer vision, as it is a prerequisite for many downstream applications, such as scene understanding [24, 27], autonomous driving [2, 17], and intelligent robotics [11]. With sufficient training annotations, prior research [3, 21, 36] has demonstrated favorable performance via deep neural networks. However, the remarkable success mostly relies on the closed-world assumption that the test data share the same underlying class label space as the trained data. Unfortunately, in many realistic scenarios, this assumption does not hold true due to the dynamic nature of real-world tasks where novel classes may emerge. Additionally, obtaining large annotated training data in open-label spaces is costly and time-consuming. To address this challenge, Open Vocabulary Object Detection (OVD) [30, 34, 35] is introduced to relax the closed-world assumption, which assumes that the novel classes extend the detection vocabulary beyond the training classes.

In recent years, a great number of efforts [1, 5, 8] have been made to improve OVD tasks. A prevalent approach in

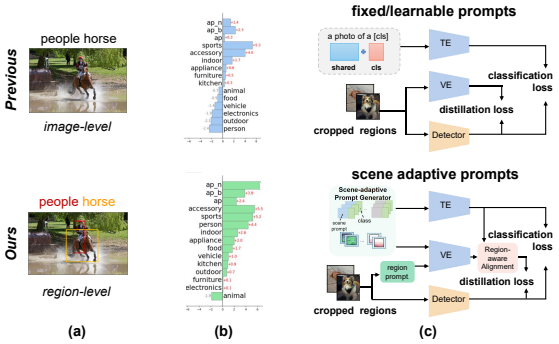

Figure 1. (a) Image-level pre-training and region-level detection. (b) The relative mAP improvement for different superclasses compared to the prompt templates on the COCO dataset. Each row corresponds to a superclass in COCO, representing distinct scenic. The bar charts on the left and right represent performance degradation and improvement, respectively. (c) Existing distillation-based OVD detectors adopt the templates or learn common prompts to adapt the knowledge of VLM. Instead, our method adopts scene-aware prompts for both visual and textual input, and incorporates the region prompts.

*Equal contribution.
†Corresponding author.

this line of research [5, 8, 30, 34] transfers the knowledge of the large-scale pre-trained visual and language models (VLM), like CLIP [25], to the detector through the distillation of feature embeddings and prompt learning. Specifically, these methods employ embeddings of image crops [5] to distill the visual feature, and replace the classifier weights with text embeddings. These text embeddings are generated by feeding prompt templates, such as "a photo of [class]", filled with base class names into the text encoder of CLIP. Nevertheless, these templates need to be constructed specifically for different tasks, and have inadequate generalization capabilities. Later, motivated by great progress in prompt tuning [38] in the NLP domain, Detpro [5] used a learnable prompt to enhance the transferred knowledge, which is concatenated with class vectors as the final prompt through learnable random vectors shared among all classes.

Despite significant development, existing distillation-based methods still show several limitations when transferring the knowledge to the detector: (i) Weak adaptability for diverse classes. Existing works fail to consider both the commonality and diversity of the class distributions when utilizing the prompt templates or sharing learnable prompts within a single modality. Figure 1 (b) illustrates the performance improvement of the learnable prompt shared by all classes on the COCO dataset compared to the prompt templates. The results reveal that some classes exhibit poor performance when learning a common prompt for all classes. (ii) Misalignment between region and image level cues. As shown in Figure 1 (a), object detection performs recognition on image regions, while the CLIP model is trained on the whole image, leading to a distribution gap that impedes detection performance. These problems result in a less efficient transfer of pre-trained visual and linguistic knowledge to the detection network.

To address the problems mentioned above, this paper introduces the Scene-adaptive and Region-aware Multi-modal Prompt (SAMP) for open vocabulary object detection, aiming to effectively integrate class-aware and region-level knowledge of VLM within the object detection framework (Figure 1 (c)). To enhance adaptability to diverse classes, we design a scene-adaptive prompt generator to construct a set of scene-specific multi-modal prompts by combining a common prompt shared across all classes with a collection of individualized prompts tailored to each scene, where the scene-specific uses insights from low-rank decomposition [16]. We then adaptively learn these scene prompts by dynamically selecting the corresponding scene prompts according to the input instance. Based on the scene-specific prompts, the more distinctive attributes associated with each class can be transferred from the pre-trained model. Furthermore, to bridge the gap between region-level object detection and image-level pre-training, we introduce a region-aware multi-modal alignment module. This module utilizes a region prompt to extract positional information from the global feature, which is then transferred to the region feature using the self-attention mechanism. Simultaneously, the textual prompt is integrated into the visual features through network mapping, enhancing the alignment of visual and language knowledge at the regional level.

In summary, this paper makes the following contributions:

- We propose scene-adaptive and region-aware multi-modal prompts for the OVD task, which is a novel paradigm to effectively adapt class-aware multi-modal knowledge from VLM to the detection network.
- The scene-adaptive prompt generator formulates a novel prompt generation and selection mechanism for both visual and text encoders, which can adaptively learn common knowledge for all classes and scene-specific knowledge to acquire better class-aware classifier weights.
- The region-aware multi-modal alignment module aligns the vision and language representation at the region level by incorporating region prompts and text prompts.
- We conduct extensive experiments on two commonly used and challenging benchmarks, OV-COCO and OV-LVIS. The proposed method consistently outperforms existing state-of-the-art methods in various settings. Combined with Faster R-CNN, our SAMP achieves 34.8% mAP_{50} of novel classes on OV-COCO and 27.8% mAP of novel classes on OV-LVIS.

2. Related Work

2.1. Open Vocabulary Object Detection

Traditional object detectors are constrained by pre-defined object classes, limiting their applicability in real-world scenarios, so recent attempts further explore Open Vocabulary Object Detection (OVD) [4, 5, 35] which can be extended to novel classes. Depending on the type of supervisory information used, OVD methods can be classified into four types: Region-aware training [1, 4, 18] is based on the cheap and abundant image-caption pairs besides the ground-truth datasets, such as OVR-CNN [35]. The Pseudo-labelling methods [37, 39] also leverage image-text pairs besides ground truths but they explicitly construct pseudo-region-text pairs to learn the correspondence in a teacher-student framework. Distillation-based methods [5, 30] distill the region embeddings from the VLM into the student model to make them compatible with text embeddings of VLM using detection. Transfer learning-based methods [15] leverage the VLM image encoder as a feature extractor, which adds negligible extra computation.

2.2. Prompt Learning

Prompt tuning [38] is a technique used in natural language processing to adapt the pre-trained VLM to the downstream

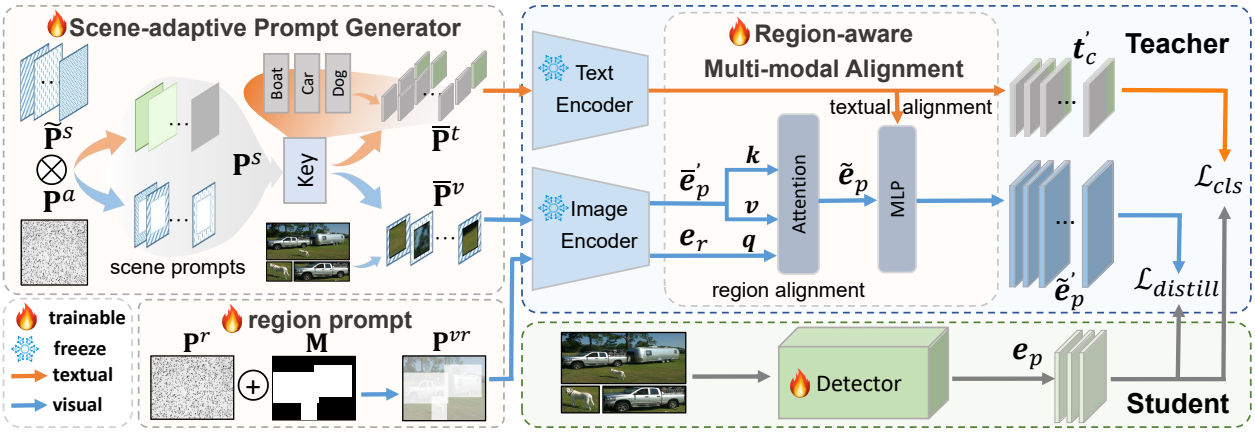

Figure 2. The framework of the SAMP model. We adopt the distillation-based framework, and insert the scene-adaptive prompt generator and region-aware multi-modal alignment module to adapt the knowledge in the pre-trained VLM to the detector.

tasks, which applies task-related textual tokens to infer task-specific knowledge. For example, the hand-crafted template "a photo of [CLASS]" in CLIP [25] is used to model the textual embedding for zero-shot prediction. However, these hand-crafted prompts have less ability to describe the downstream task because they do not consider the specific knowledge of the current task. To address the above problem, Context Optimization(CoOp) [38] replaces the hand-crafted prompts with the learnable soft prompts inferred from the labeled few-shot samples. Visual prompt learning (VPT) [12] tunes embedded visual prompts with a frozen pre-trained ViT backbone supervised by downstream objectives, which achieves better transfer. MaPLe [13] proposed multi-modal prompt learning to improve alignment between the vision and language representations.

3. Preliminaries

3.1. Problem Definition

Open Vocabulary Object Detection (OVD) aims to detect objects from novel classes beyond the base classes on which the detector is trained. In the training stage, we have base classes $\mathcal{C}^B$ and datasets $\mathcal{D}^{tr} = \{\mathbf{I}_i, \mathcal{O}_i\}_{i=1}^{|\mathcal{D}^{tr}|}$, where $\mathbf{I}_i$ is the i-th image and $\mathcal{O}_i = \{o_{ij}\}_{j=1}^{|\mathcal{O}_i|}$ are the labeled instances in the image, and each object o_{ij} consists of the class label $y \in \mathcal{C}^B$ and bounding box. In the testing stage, there are novel classes $\mathcal{C}^N$ that need to be detected and $\mathcal{C}^B \cap \mathcal{C}^N = \emptyset$.

3.2. Distillation-based OVD Method

As illustrated in Figure 2, the knowledge distillation-based OVD paradigm as [8, 30] employs a teacher-student framework to transfer the knowledge pre-trained on large-scale data to the detector.

The teacher network adopts the pre-trained CLIP model, with visual and textual encoders denoted as $\mathcal{V}$ and $\mathcal{T}$, respectively. For the student network, we employ the two-stage Faster-RCNN network [26]. The backbone processes an input image $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ to generate feature maps $\mathbf{F}$, and RPN generates a set of proposals $\mathcal{P} \in \mathbb{R}^4$. Subsequently, the R-CNN head performs RoI Align on $\mathbf{F}$ to extract proposal embeddings $\mathcal{E}_o = \{\boldsymbol{e}_p\}_{p\in\mathcal{P}} \in \mathbb{R}^d$ for distillation and detection. To enable the detection of both trained base classes and emerging novel classes, a common operation is to replace the fixed classifiers with text embedding representations of each class $\boldsymbol{t}_c \in \mathbb{R}^d$ generated by $\mathcal{T}$. These text embeddings $\boldsymbol{t}_c$, obtained by processing the prompts corresponding to each class with a template of "a photo of [class]", can be directly extended to deal with the emerging classes at test time. The classification probability for each proposal is expressed as:

$$P_{\mathcal{C}}(p,c) = \frac{\exp(\frac{\boldsymbol{e}_p \cdot \boldsymbol{t}_c}{||\boldsymbol{e}_p|| \cdot ||\boldsymbol{t}_c||})}{\sum_{c' \in \mathcal{C}^B \cup \mathcal{C}^N \cup \{\text{bg}\}} \exp(\frac{\boldsymbol{e}_p \cdot \boldsymbol{t}_{c'}}{||\boldsymbol{e}_p|| \cdot ||\boldsymbol{t}_{c'}||})}, \quad (1)$$

where $\cdot$ is the dot product, $\{\text{bg}\}$ is the background class, which is a trainable embedding $\boldsymbol{t}_{bg} \in \mathbb{R}^d$.

Also, the proposal embeddings $\bar{\mathcal{E}}_o = \{\bar{\boldsymbol{e}}_p\}_{p\in\mathcal{P}} \in \mathbb{R}^d$ in the teacher network are obtained by the visual coder $\mathcal{V}$ for aligning region embeddings with the student embeddings.

3.3. Prompt Tuning

In addition to the prompt templates, prior research [38] has utilized prompt tuning to incorporate learnable prompts into the textual encoder to transfer pre-trained model knowledge effectively. Specifically, for a given class c, the concatenated tokens of learnable prompts $\mathbf{P}^t$ and the class name are fed into the text encoder $\mathcal{T}$ to obtain the classifier weights $\boldsymbol{t}_c = \mathcal{T}([\mathbf{P}^t; \boldsymbol{w}_1, ..., \boldsymbol{w}_{L_1}; \text{cls}])$ for the detector. $\{\boldsymbol{w}_1, ..., \boldsymbol{w}_{L_1}\}$ are the tokens of the name of class c, and L_1 is the total sequence length, [cls] is the class token, $[;]$ represents concatenation along the token length dimension.

For visual prompts, the common practice is also prepending the prompts $\mathbf{P}^v$ with the input image into the layers of the visual encoder. The input embedding layer transforms the input image into a sequence of patch embeddings, and the concatenated tokens are $[\mathbf{P}^v; \boldsymbol{v}_1, ..., \boldsymbol{v}_{L_2}; \text{cls}]$. where $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_{L_2}\}$ is the set of image tokens and [cls] is the class token, L_2 is the length of the patch tokens.

4. Method

Despite greater advances in distillation-based methods, there remain two challenges when transferring the knowledge from the VLM to the detector: (1) existing methods only learn a common prompt for all classes, which shows weak adaptability for diverse classes. (2) object detection performs recognition on image regions, while the CLIP model is trained on the whole image, leading to a distribution gap that hampers classification performance. To address the aforementioned challenges, we propose an OVD framework including a Scene Adaptive Prompt generator(SAP) and Region-aware Multi-modal Alignment (RMA) module, as shown in Figure 2. The SAP can generate class-aware prompts for both text and visual input, which can adaptively learn common knowledge for all classes and acquire better class-aware knowledge. The RMA module incorporates the region and semantic prompt to modulate the region features for better generalizable region embeddings and to narrow the distribution gap.

4.1. Scene-adaptive Prompt Generator

Transferring pre-trained VLM knowledge to the detection network requires accurate prompts for higher performance. However, prompt learning approaches that employ the same generation method for all classes or pre-defined prompt templates would lead to a strong inductive bias. Therefore, we deploy a scene-adaptive prompt generator to transfer the class-specific pre-trained VLM knowledge. The generator classifies classes into different scenes and applies different prompts to each scene, then these scene prompts are adaptively learned through the selection mechanism.

Constructing scene-adaptive prompts. Instead of manually dividing each class into different scenes, our scene-adaptive prompt generator employs an innovative mechanism to generate scene-adaptive prompts suitable for both visual and textual inputs. This mechanism comprises two primary components: a common prompt designed to acquire knowledge shared among all classes, and a set of scene prompts constructed by low-rank decomposition to acquire scene-specific knowledge.

Specifically, taking the textual branch as an example, we define $\mathbf{P}^a \in \mathbb{R}^{L\times d}$ as the common prompt to learn the knowledge for all classes, where L denotes the prompt length and d corresponds to the feature embedding size. Additionally, we construct a series of prompts tailored to specific scenes, denoted as $\{\tilde{\mathbf{P}}_i^s\}_{i=1}^m$, where each prompt $\tilde{\mathbf{P}}_i^s \in \mathbb{R}^{L\times d}$ is dedicated to learning scene-specific knowledge. Then the scene prompt $\mathbf{P}_i^s \in \mathbb{R}^{L\times d}$ can be obtained as follows:

$$\mathbf{P}_i^s = \mathbf{P}^a \otimes \tilde{\mathbf{P}}_i^s, \tag{2}$$

where $\otimes$ is the Hadamard product.

To facilitate more efficient and effective computations, inspired by the effective low-rank methods discussed in [16], which have demonstrated robust performance in optimizing tasks within a low-rank subspace, we propose to parameterize $\tilde{\mathbf{P}}_i^s$ as a low-rank matrix. This matrix is represented as the product of two low-rank vectors: $\mathbf{u}_i^s \in \mathbb{R}^L$ and $\mathbf{v}_i^s \in \mathbb{R}^d$:

$$\mathbf{P}_i^s = \mathbf{P}^a \otimes (\mathbf{u}_i^s \cdot (\mathbf{v}_i^s)^\top). \tag{3}$$

Decomposing scene prompts into rank-one subspaces serves to enhance the effective encoding of scene-specific information within the model. By utilizing the Hadamard product, this approach empowers the model to efficiently leverage both common and scene-aware knowledge for accurate predictions.

Selecting mechanism. To adaptively learn the scene prompts, we develop a selection mechanism to dynamically choose corresponding scene prompts to input instances. In detail, each prompt is associated as a value to a learnable key: $\{(k_1, \mathbf{P}_1^s), (k_2, \mathbf{P}_2^s), ..., (k_M, \mathbf{P}_m^s)\}$, where $k_i \in \mathbb{R}^d$ is obtained through a mapping function $k_i = g(\mathbf{P}_i^s)$. The function g can be implemented as a Multi-Layer Perceptron (MLP) layer. Given a proposal p belongs to the class c, we compute the similarity between the feature embedding $\bar{\mathbf{e}}_p$ and the keys:

$$\alpha_i = \gamma(\bar{\mathbf{e}}_p, k_i), \tag{4}$$

where γ represents the cosine similarity function, and α is a weighting vector that determines the contribution of each prompt. Subsequently, we select the set of the top-n most similar keys $\mathcal{K}_p$ based on alpha α and combine the prompts to represent the final selected prompts:

$$\bar{\mathbf{P}}_c^t = \sum_{k_i \in \mathcal{K}_p} \alpha_i \mathbf{P}_i^s. \tag{5}$$

According to select prompts in an instance-wise fashion, we can adaptively aggregate classes with similar features to update the scene-aware prompt, thus being able to consider both the commonality and specificity of the classes.

Generating multi-modal prompts. In the textual branch, leveraging the scene-adaptive generator and the selection mechanism, we can obtain scene-specific prompts $\bar{\mathbf{P}}_c^t$ for class c to adaptively transfer the knowledge from the pre-trained model. The input embedding is formulated as:

$$\boldsymbol{t}_c' = \mathcal{T}([\bar{\mathbf{P}}_c^t; \boldsymbol{w}_1, ..., \boldsymbol{w}_{L_1}; \text{cls}]). \tag{6}$$

As the classification loss is updated, the network gradually learns shared and scene-specific knowledge by updating $\mathbf{P}^a$, $\mathbf{u}_i^s$, and $\mathbf{v}_i^s$.

Also, for the visual prompts, we can obtain:

$$\bar{\mathbf{e}}_p' = \mathcal{V}([\bar{\mathbf{P}}_p^v; \boldsymbol{v}_1, ..., \boldsymbol{v}_{L_2}; \text{cls}]), \tag{7}$$

where $\bar{\mathbf{P}}_p^v$ is selected scene prompt for the $\bar{\mathbf{e}}_p$. The region features $\bar{\mathbf{e}}_p'$ are used for distillation with the detector in the training stage.

4.2. Region-aware Multi-modal Alignment

While the scene-adaptive prompts can improve the knowledge transfer of the pre-trained model, the gap between image-level pre-training and region-level detection remains unresolved and there is no clear alignment between visual and textual space. To tackle this issue, we propose a region-aware multi-modal alignment module that incorporates positional information from global features into region features by region prompts, and incorporates textual insights into visual features, which helps to align visual and textual knowledge at the region level.

Constructing region-aware prompt. As the global feature context inherently includes positional information, we introduce a learnable region prompt coupled with the region mask of each image to interact with the tokens across the entire image, which extracts positional information and subsequently transfers it to the region features.

Specifically, we first introduce a learnable region prompt $\mathbf{P}^r \in \mathbb{R}^{H\times W\times 3}$ consistent with the dimensions of the input image. Then, we construct a position mask $\mathbf{M} \in \mathbb{R}^{H\times W\times 3}$ for each image, in which the pixels within the ground truth boxes are assigned a value 1, while the remaining regions are set to 0. The region prompt is conducted as $\mathbf{P}^{vr} = \mathbf{P}^r \oplus \mathbf{M}$, where $\oplus$ is the pixel-wise addition operator.

Then we combine the region prompt with patch tokens of the whole image $\mathbf{X}_g$ to extract the position representation from the CLIP:

$$\boldsymbol{e}_r = \mathcal{V}([\mathbf{X}_g; \mathbf{P}^{vr}; \mathbf{x}_{[r]}]), \tag{8}$$

where $\mathbf{x}_{[r]}$ plays the same role as the [cls] token, capturing positional information based on the interaction between the region prompt and other tokens, and $[;]$ is the concatenation operation.

To incorporate the position information, we integrate the embedding of the region prompt with the object feature $\bar{\boldsymbol{e}}_p$. This fusion is achieved through the use of multi-head self-attention modules (MSA) to enable effective interaction between the region embedding and object features. MSA first maps each feature into three vectors, $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v} \in \mathbb{R}^{d_h}$, with linear projection parameterized by $\mathbf{W}_q$ and $\mathbf{W}_{kv}$, i.e.:

$$[\boldsymbol{k}, \boldsymbol{v}] = \bar{\boldsymbol{e}}_p' \mathbf{W}_{kv} \quad \boldsymbol{q} = \boldsymbol{e}_r \mathbf{W}_q, \tag{9}$$

$$\mathbf{A} = \text{softmax}(\boldsymbol{q}\boldsymbol{k}^\top / {d_h}^{1/2}). \tag{10}$$

The attention weights $\mathbf{A}$ are used to choose and aggregate information from region prompts. The final output is obtained by concatenating the object feature with the attention value:

$$\tilde{\boldsymbol{e}}_p = (\mathbf{A}\boldsymbol{v}^\top)\mathbf{W}_o + \bar{\boldsymbol{e}}_p, \tag{11}$$

where $\mathbf{W}_o$ denotes the parameterized linear projection.

Aligning multiple modalities. After incorporating position information, we further integrate text prompts into the visual features at the output end. This integration aims to enhance the transfer of information across multiple modalities, facilitating the accurate computation of the classifier's visual features and the similarity with text embeddings.

For a region proposal p belongs to an image $\mathbf{I}_j$, if $\mathbf{I}_j$ contains ground truth information of $\mathcal{C}_j$ multiple classes, we compute the mean value of the textual embeddings corresponding to each class $\hat{\boldsymbol{t}}_j'$. Then an interaction mechanism that allows modulating and augmenting semantic prompts to the visual features is proposed:

$$\hat{\boldsymbol{t}}_j' = \frac{1}{|\mathcal{C}_j|} \sum_{c_j \in \mathcal{C}_j} \boldsymbol{t}_{c_j}', \tag{12}$$

$$\tilde{\boldsymbol{e}}_p' = f([\tilde{\boldsymbol{e}}_p, \hat{\boldsymbol{t}}_j']), \tag{13}$$

where f is a 2-layer MLP module with the sigmoid activation function.

By incorporating the positional and textual information in an ensemble approach, the visual prompts and the region prompt can be dynamically learned, and the region feature $\tilde{\boldsymbol{e}}_p'$ can enjoy stronger transferability for better distillation.

4.3. Overall Training and Inference Process

In the training stage, we generate text embeddings $\boldsymbol{t}_c'$ as the classifier weight for class c, and calculate the similarity $P_{\mathcal{C}}(p, y_p)$ with the visual embedddings $\mathcal{E}_o = \{\boldsymbol{e}_p\}_{p\in\mathcal{P}}$ of the student network. The classification loss is:

$$\mathcal{L}_{cls} = -\sum_{p\in\mathcal{P}} \log P_{\mathcal{C}}(p, y_p). \tag{14}$$

In the process of feature distillation, as in previous methods [8, 30], we obtain the final proposal embeddings $\tilde{\mathcal{E}}' = \{\tilde{\boldsymbol{e}}_p'\}_{p\in\mathcal{P}}$ through the teacher network for aligning region embeddings with the student embeddings: $\mathcal{L}_O = \mathcal{L}_1(\tilde{\mathcal{E}}', \mathcal{E}_o)$. $\mathcal{L}_1$ is the Mean Absolute Error. We also distill the features of the whole image, and the distillation loss denotes as $\mathcal{L}_G$, so $\mathcal{L}_{distill}$ consists of $\mathcal{L}_O$ and $\mathcal{L}_G$. The total training loss is:

$$\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_{cls} + \mathcal{L}_{rpn} + \mathcal{L}_{reg}, \tag{15}$$

where $\mathcal{L}_{rpn}$ denotes the classification and regression losses of the RPN, and $\mathcal{L}_{reg}$ is the regression loss of the detector.

Method	Supervision	Backbone	Detector	Prompt	mAP_{50}^{N}	mAP_{50}^{B}	mAP_{50}
Region-Aware Pretraining							
OV-RCNN [35]	Caption	R50-C4	FRCNN	-	22.8	46.0	39.9
LocOv [1]	Caption	R50-C4	FRCNN	-	28.6	51.3	45.7
MEDet [4]	Caption	R50-C4	FRCNN	T(cat)	32.6	53.5	48.0
VLDet [18]	Caption	R50-C4	FRCNN	T(cat)	32.0	50.6	45.8
RO-ViT [14]	ALIGN	ViT-B/16	MRCNN	T(cat)	30.2	-	41.5
Pseudo-Labelling							
RegionCLIP [37]	Caption	R50-C4	FRCNN	T (cat)	26.8	54.8	47.5
Detic [39]	Caption	R50-C4	FRCNN	T (cat)	27.8	47.1	45.0
PromptDet [6]	LAION-novel	R50-FPN	MRCNN	L (cat+desc)	26.6	-	50.6
PB-OVD[7]	Caption	R50-C4	FRCNN	T (cat)	29.1	44.4	40.4
XPM [10]	Caption	R50-C4	FRCNN	-	27.0	46.3	41.2
GOAT [29]	Caption	R50	FRCNN	T (cat)	31.7	51.3	46.1
Transfer Learning-based							
F-VLM [15]	-	R50-FPN	MRCNN	T (cat)	28.0	-	39.6
Knowledge Distillation-based							
ViLD [5]	CLIP	R50-FPN	FRCNN	T (cat)	27.6	59.5	51.2
ZSD-YOLO [33]	-	CSP-DN53	YOLOv5x	T (cat+desc)	13.6	31.7	19.0
HierKD [22]	Caption	R50-FPN	ATSS	T (cat/desc)	20.3	51.3	43.2
OADP [30]	-	R50-FPN (SoCo)	FRCNN	T (cat)	30.0	53.3	47.2
BARON [32]	CLIP	R50-FPN (SoCo)	FRCNN	T (cat)	34.0	60.4	53.5
Ours	CLIP	R50-FPN (SoCo)	FRCNN	SAP	**34.8**	**61.4**	**54.2**

Table 1. The comparision results with other methods on the OV-COCO dataset.

In the testing phase, we only utilize the trained student network to extract features. The classifiers can be extended to novel classes by incorporating the names of newly introduced classes and selecting the trained scene prompts dynamically.

5. Experiments

In this section, we comprehensively evaluate our SAMP on the OVD task. More results are provided in the Appendix.

5.1. Datasets and Metrics

Datasets. We evaluate our method on two commonly-used datasets OV-COCO and OV-LVIS. Following the benchmark proposed in [35], the classes in the COCO dataset [19] are divided into 48 base classes and 17 novel classes. The model is trained on the 48 base classes, and then evaluated on the validation set containing both the base and the novel classes. We also conduct experiments on the LVIS v1.0 [9] dataset. On the LVIS dataset, the model is trained on 461 common classes and 405 frequent classes, and subsequently evaluated using the LVIS validation set.

Metrics. We evaluate the performance under the "generalized" setting, where the model needs to predict both base and novel classes for completeness. Consistent with prior works [35], for OV-COCO, we report the mAP_{50} for base, novel, and all classes, denoted as $mAP_{50}^{B}, mAP_{50}^{N}, mAP_{50}$. For OV-LVIS, we report AP_r, AP_c, AP_f, and AP for rare (novel), common, frequent, and all categories.

5.2. Experimental Details

Following previous works [5, 30], our model utilizes the Faster-RCNN with a ResNet-50 backbone, and initializes the student backbone using SoCo [12]. We set the batch size as 16 and run the SGD optimizer with an initial learning rate of 0.02 and a weight decay of 2.5×10^{-5}. For the teacher network, we adopt the ViT-B/32 CLIP. Under the OV-COCO setting, we train the detector for 40, 000 iterations. For OV-LVIS, we use a 2x training schedule. We set 15 scene prompts on the OV-COCO dataset and 100 on OV-LVIS, and select 5 most prompts for each instance.

5.3. Benchmark Results

Results on OV-COCO. To evaluate the effectiveness of our approach, we compare the results with previous state-of-the-art methods on the COCO dataset in Table 1, and we mainly compare with the RCNN-based detectors. As in [40], we classify the OVD methods into four types based on the availability of supervisory information, such as image-caption pairs, pseudo pairs, or pre-trained models. T(cat) denotes they adopt the prompt templates for text input, and L(cat) means they adopt the learnable prompts, "desc" is class descriptions obtained from WordNet [23].

Compared to previous knowledge distillation-based models, our method has demonstrated a 7.2% improvement over the ViLD method, which utilizes hand-crafted tem-

Method	Backbone	Detector	Teacher	Prompts	AP_r	AP_c	AP_f	AP
Detic [39]	R50-FPN	MRCNN	-	T(cat)	17.8	26.3	31.6	26.8
PromptDet [6]	R50-FPN	MRCNN	-	L(cat+desc)	21.4	23.3	29.3	25.3
CondHead [31]	R50-C4	MRCNN	-	T(cat)	19.9	28.6	35.2	29.7
ViLD-ens [5]	R50-FPN	MRCNN	CLIP	T(cat)	16.7	26.5	34.2	27.8
DetPro [5]	R50-FPN (SoCo)	MRCNN	CLIP	L(cat)	20.8	27.8	32.4	28.4
F-VLM [15]	R50-FPN	MRCNN	-	T(cat)	18.6	-	-	24.2
OADP [30]	R50-FPN (SoCo)	FRCNN	CLIP	T(cat)	21.9	28.4	32.0	28.7
BARON [32]	R50-FPN (SoCo)	FRCNN	CLIP	L(cat)	23.2	29.3	32.5	29.5
Ours	R50-FPN (SoCo)	FRCNN	CLIP	SAP	**27.8**	**30.4**	**36.8**	**34.3**

Table 2. The comparision results on LVIS dataset.

method	backbone	detector	VOC	COCO	
			mAP_{50}	mAP	mAP_{50}
ViLD	R50-FPN	MRCNN	72.2	36.6	55.6
OV-DETR	R50-C4	Def-DETR	76.1	38.1	58.4
DetPro	R50-FPN (SoCo)	MRCNN	74.6	34.9	53.8
F-VLM	R50-FPN	MRCNN	-	32.5	53.1
GridCLIP	CLIP (R50)	FCOS	70.9	34.7	52.2
Ours	R50-FPN(SoCo)	MRCNN	**76.8**	**38.6**	**58.9**

Table 3. OVD performance under the CDTE protocol on the test set of Pascal VOC and validation set of COCO.

Method	mAP^B_{50}	mAP^N_{50}	mAP_{50}
Baseline	56.5	27.6	48.2
Baseline+SAP	59.6	32.8	53.1
Baseline+RMA	59.3	32.3	52.7
Ours	**61.4**	**34.8**	**54.2**

Table 4. The effectiveness of each component.

	shared	scene	mAP^B_{50}	mAP^N_{50}	mAP_{50}
textual	√		57.1	29.3	48.9
		√	58.8	31.6	50.8
visual	√		57.3	28.8	49.3
		√	58.5	30.7	50.8
multi-modal	√		57.8	29.1	49.8
		√	**59.6**	**32.8**	**53.1**

Table 5. The effectiveness of of different prompts in generator.

plates. BARON performs well on mAP^N_{50}, but is lower on 6.6% on mAP^B_{50} and 5.1% on mAP_{50}. Therefore our method reduces the performance gap in known classes and improves over the novel classes as compared to all previous works. We can see that region-aware pre-training methods generally achieve better results because they can utilize information from a large number of image-text pairs to pre-train for the detection task, while our method just adapts the CLIP to detectors.

Results on OV-LVIS. We also conduct experiments on the OV-LVIS dataset as shown in Table 2. We mainly compare with the RCNN-based methods, and other results can

Method	mAP^B_{50}	mAP^N_{50}	mAP_{50}
Baseline	56.5	27.6	48.2
+random	57.1	28.4	49.7
+decomposition	**58.8**	**31.6**	**50.8**
+region prompt	58.7	30.4	49.7
+text prompt	57.9	29.6	50.2
+all	**59.3**	**32.3**	**52.7**

Table 6. The effectiveness of low-rank decomposition and fusing different prompts.

be seen in the Appendix. Our best model achieves 27.8% AP_r, which significantly outperforms the best existing approach by 4.6 points. The PromptDet and DetPro adopt the learnable prompts for text embedding, and our method can outperform them by 6.4% and 7.0% respectively.

Cross-dataset Transfer Evaluation (CDTE). To evaluate the generalization ability of SAMP, we also perform cross-dataset transfer evaluation, in which the model is trained on the LVIS base classes and evaluated on the OV-COCO and Pascal-VOC datasets. We replace the LVIS with other datasets' vocabulary embeddings to perform the transfer detection without finetuning. The results can be seen in Table 3. It can be seen that the model trained on LVIS has better migration and generalization capabilities. After analysis, it is found that the categories in LVIS are relatively fine-grained categories, and there are many similar categories on COCO and VOC datasets.

5.4. Analytical Experiment

Ablation of different component. Table 4 shows the effectiveness of each module in our SAMP framework. The baseline method is the re-implemented ViLD-ensemble as in [30], where the classifier weights are hand-crafted prompts. The second and third rows in the table show that the proposed scene-adaptive prompt generator and the location-guided multi-modal fusion all benefit the adaptation performance. Applying the scene adaptive prompt generator module brings a 5.2% mAP^N_{50} gain, suggesting that

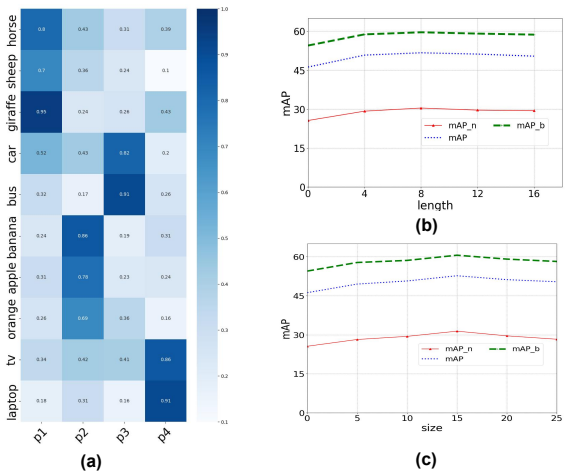

Figure 3. (a) The similarity between the learned prompts and the classes. (b),(c) The results of different prompt lengths and different prompt sizes on OV-COCO dataset.

it can better adapt the pre-trained knowledge to the detector, while the region level multi-modal fusion brings 4.7% mAP_{50}^{N} gain, meaning that align the visual and language is better for adaptation. The combination of generator and fusion can further improve the performance.

Ablation in each component. In Table 5, we show the results with only shared prompts and adaptive prompts for the text branch, visual branch, and the whole method. The shared prompt means that all classes learn a common prompt. It can be observed that all learnable prompts enhance the performance of the base classes. Furthermore, it is evident that both class-aware visual and text prompts contribute to boosting performance in novel classes by 1.9% and 2.3% respectively. Moreover, we compared the effect of the low-rank decomposition in the upper part of Table 6, and we can see that there is a better improvement on both the base and novel classes when constructing different scene prompts using the low-rank decomposition.

Table 6 also shows the results of incorporating different prompts in the RMA module. We can observe that both region and semantic prompts significantly enhance performance, leading to accuracy improvements across all classes by 1.5% and 2.0%, respectively. Region prompts improved by 2.2% on base classes, while only text prompt is slightly lower, perhaps because it can better fit base classes with region prompts to distill the region information. Furthermore, the accuracy of all classes is further improved by combining them. These results indicate that the proposed semantic and region fusion is an effective approach for detection.

Analysis of the scene prompt. Figure 3 (a) shows the similarity between the learned textual prompts and some classes, p1-p4 represents the learned prompts. We can see that prompt 1 is more similar to the animal classes while prompt 2 is more similar to classes of fruit. Figure 3 (b) shows the results for different prompt lengths on the COCO dataset. To evaluate the effect of the prompt length, we study 4, 8, 12, and 16 prompt tokens. It can be observed that the differences are fairly small whereas the context length 8 can achieve a better performance. Figure 3 (c) summarizes the results for different sizes of scene prompts. We can observe that the best size for the OV-COCO dataset is 15.

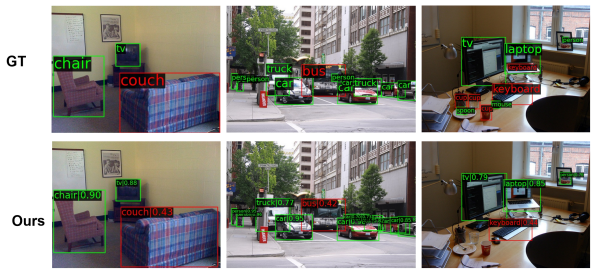

Figure 4. Visualization of the detection results.

5.5. Visualization

In Figure 4, we visualize the detection results of our method. The top row displays the ground truth labels of the objects, while the bottom row shows the predictions of our SAMP model. This visualization effectively demonstrates the capabilities of our SAMP model. In particular, our model excels at detecting objects from both novel and base classes. For instance, in the first column, our model accurately distinguishes between a couch and a chair.

6. Conclusion

In this work, we propose scene-adaptive and region-aware multi-modal prompts for the OVD task, which enables adaptive migration of class-aware knowledge and helps to narrow the distribution gap between detection and pre-training. Unlike previous works that rely on prompt templates or learn a common prompt for all classes, we formulate a novel prompt generation mechanism to construct a set of scene prompts, and design an instance-based selection mechanism to adaptively learn the prompts that are suitable for visual and textual inputs. Furthermore, we incorporate the region prompt and the semantic prompt into region features to align visual and linguistic representations at the region level. Experiments show that our method can better transfer the knowledge for both base classes and novel classes. We hope that our work can help other researchers gain better insight into the OVD problem and develop better open vocabulary detectors.

Acknowledgements

This work is supported by grants No.KZ46009501.

References

[1] Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. In *DAGM German Conference on Pattern Recognition*, pages 393–408. Springer, 2022. 1, 2, 6

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[4] Peixian Chen, Kekai Sheng, Mengdan Zhang, Mingbao Lin, Yunhang Shen, Shaohui Lin, Bo Ren, and Ke Li. Open vocabulary object detection with proposal mining and prediction equalization. *arXiv preprint arXiv:2206.11134*, 2022. 2, 6

[5] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1, 2, 6, 7

[6] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European Conference on Computer Vision*, pages 701–717. Springer, 2022. 6, 7

[7] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision*, pages 266–282. Springer, 2022. 6

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. 1, 2, 3, 5

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6

[10] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 6

[11] Minyoung Hwang, Jaeyeon Jeong, Minsoo Kim, Yoonseon Oh, and Songhwai Oh. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6683–6693, 2023. 1

[12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3, 6

[13] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3

[14] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11144–11154, 2023. 6

[15] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *The Eleventh International Conference on Learning Representations, ICLR*, 2023. 2, 6, 7

[16] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *6th International Conference on Learning Representations, ICLR*, 2018. 2, 4

[17] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1

[18] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *The Eleventh International Conference on Learning Representations, ICLR*, 2023. 2, 6

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[20] Xianglong Liu, Shihao Bai, Shan An, Shuo Wang, Wei Liu, Xiaowei Zhao, and Yuqing Ma. A meaningful learning method for zero-shot semantic segmentation. *Science China Information Sciences*, 66(11):210103, 2023. 1

[21] Yuqing Ma, Hainan Li, Zhange Zhang, Jinyang Guo, Shanghang Zhang, Ruihao Gong, and Xianglong Liu. Annealing-based label-transfer learning for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11454–11463, 2023. 1

[22] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022. 6

[23] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 6

[24] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al.

Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 1

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 3

[27] Jiakai Wang, Aishan Liu, Xiao Bai, and Xianglong Liu. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. *IEEE Transactions on Image Processing*, 31:598–611, 2021. 1

[28] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8565–8574, 2021. 1

[29] Jiong Wang, Huiming Zhang, Haiwen Hong, Xuan Jin, Yuan He, Hui Xue, and Zhou Zhao. Open-vocabulary object detection with an open corpus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6759–6769, 2023. 6

[30] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2023. 1, 2, 3, 5, 6, 7

[31] Tao Wang. Learning to detect and segment for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7051–7060, 2023. 7

[32] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 6, 7

[33] Johnathan Xie and Shuai Zheng. Zero-shot object detection through vision-language embedding alignment. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1–15. IEEE, 2022. 6

[34] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 1, 2

[35] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1, 2, 6

[36] Xiaowei Zhao, Yuqing Ma, Duorui Wang, Yifan Shen, Yixuan Qiao, and Xianglong Liu. Revisiting open world object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[37] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2, 6

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 3

[39] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 2, 6, 7

[40] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *arXiv preprint arXiv:2307.09220*, 2023. 6